DETECTING THE VIOLATION OF HOMOGENEITY IN MIXED MODELS: A CASE STUDY

FANG XICHENG

NATIONAL UNIVERSITY OF SINGAPORE

2013

DETECTING THE VIOLATION OF HOMOGENEITY IN MIXED MODELS: A CASE STUDY

FANG XICHENG

(B.Sc. Nanyang Technological University)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2013

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my supervisor Professor Li Jia-Liang, who is a great mentor not only in academic but also in daily life. I would like to thank him for his guidance, encouragement, time, and patience through the learning process of this thesis. Next, I would like to thank all my seniors and classmates for discussion on various topics in research. I also thank all my friends who have supported me both by keeping me harmonious and helping me to make life easier. I wish to express my gratitude to the university and the department for supporting me through NUS Graduate Research Scholarship. Finally, I will thank my family for their love and support.

CONTENTS

SUMMARY

There has been no systematic approach for checking the homogeneity assumption for generalized linear mixed-effects models. Extreme outliers that behave differently from the population may cause problems for model fitting and interpretation. We propose two tests based on random effects where the covariance matrices may be computed from the fitted model covariance parameters or the empirical variation of random effects. The tests may serve as a tool to detect outliers that violate homogeneity in mixed-effects models. Extensive simulations are carried out to assess the performance of our methods. A real case study of arthritis disease is included to provide further illustration. The results suggest removing outliers may change the signs and magnitude of important predictors in the model.

LIST Of NOTATIONS

M^T	transpose of a matrix M
$vec(A)$	vectorization of matrix A, converts the $m \times n$ matrix A into a mn vector by stacking the columns of the matrix A on top of one another
$vech(A)$	half-vectorization of symmetric matrix A, vectorizing the lower triangular part of $n \times n$ matrix A into a $n(n+1)/2 \times 1$ column vector
$\otimes$	tensor product

List of Tables

CHAPTER 1

Introduction

The standard linear model and ordinary least squares regression are well known and widely used in the real world application. But they are generally inappropriate for dependent variables. Dependent data raises in many contexts, the two most common of which are hierarchical data and longitudinal data. A hierarchical data model is a data model in which the data are sampled from two or more levels or non-nested multilevel data. In hierarchical date model, the data is organized into a tree like structure. A typical example is the parent-child relationships: each parent may have many children, but each child has only one parent. While longitudinal study is a correlational research study that involves repeated observations of the same

individuals or variables over long time periods and longitudinal data are collected over time. In all these cases, it is not reasonable to assume the observations within the same higher level unit or observations from the same person to be independent. In contrast, mixed effect model make it possible to take account the dependencies.

Linear mixed models (Laird and Ware (1982)) assume two types of variation, within cluster and between clusters. And two types of coefficients are distinguished: population averaged and cluster specific, while the latter are random and estimated as posteriori means. The mixed model technique is a combination of the frequentist and Bayesian approaches, which provides a powerful and flexible tool for the analysis of grouped data arising in many areas including agriculture, biology, economics, manufacturing, and geophysics and offers the flexibility in modeling the within group correlation often presented in grouped data by handling balanced and unbalanced data in a unified framework. Mixed effect models became more and more popular in the recent decades.

Mixed-effects models are widely used in longitudinal studies and most longitudinal studies are observational. In real life application, the assignment of treatment may be beyond the control of the investigator and randomized experiment cannot be carried out for a variety of reasons: a randomized experiment would violate ethical standards or may be impractical, the investigator may lack the requisite influence. In this case, an observational study (Rosenbaum (2002)) which draws

inferences about the possible effect of a treatment on subjects where the assignment of subjects is outside the control of the investigator may be implemented. Although observational studies cannot be used as reliable sources to make statements of fact about the safety, efficacy or effectiveness, they can still be useful for some other things: provide information on real world practise and detect signals in the general population, help formulate hypotheses to be tested in subsequent experiments and provide information for clinical practise. Compared to controlled studies, observational studies typically have a much larger data set and can avoid the ethical dilemma that of taking away the right of the participant to make his or her own decisions. For example, all the studies on the harm of smoking are based on observational studies.

An observational study is called longitudinal if it includes multiple observations for each individual over time. Longitudinal studies can measure changes and give greater validity to correlations observed by making multiple measurements over time, sometimes several decades. They often follow specific subgroups of a population, and can be built to keep track of data from specific individuals. Data is first collected at the beginning of the study and may then be gathered repeatedly throughout the length of the study. In this way, confounders and bias can be accounted for, which reduces erroneous conclusions found in the surveys. Longitudinal studies allow researchers to look at changes over time. Because of this,

longitudinal methods are particularly useful when studying development and lifespan issues. However, the amount of time required and high cost are the drawbacks of longitudinal research. To fix ideas, we consider the setting of longitudinal data in this thesis but the results can be readily applied to spatially dependent data or cluster data.

Longitudinal data may be unbalanced because of patients death and absent. Due to the unbalanced nature, many data sets cannot be analyzed using multivariate regression techniques. But a natural alternative is that the subject specific profiles can often be well approximated by linear regression functions. Let the random variable Y_{ij} denoted the response of interest for the ith cluster $(i = 1, \cdots, N; j = 1, \cdots, n_i)$ and $\boldsymbol{y}_i$ denote the column vector of $Y_{i1}, \cdots, Y_{in_i}$. As what we do in the linear regression model, we can assume a first stage model

$$\boldsymbol{y}_i = \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i,$$

where $\boldsymbol{Z}_i$ is a $n_i \times q$ matrix of known covariates and $\boldsymbol{\beta}_i$ is a q-dimensional vector of unknown subject specific regression coefficients. $\boldsymbol{\varepsilon}_i$ is a vector of residual components $\varepsilon_{ij}, j = 1, \cdots, n_i$.

In a second step, a multivariate regression model of the form

$$\boldsymbol{\beta}_i = \boldsymbol{K}_i\boldsymbol{\beta} + \boldsymbol{b}_i,$$

is used to explain the observed variability between the subjects with respect to the subject specific regression coefficients $\boldsymbol{\beta}_i$. $\boldsymbol{K}_i$ is a $q \times p$ matrix of known covariates,

$\boldsymbol{\beta}$ is a p dimensional vector of unknown regression parameters and $\boldsymbol{b}_i$ are assumed to be independent q dimensional multinormal distribution with zero mean vector (Verbeke and Molenberghs (2000)).

If we combine the two-stage model together, we come to the Linear Mixed Effects (LME) model originally introduced by Laird and Ware (1982). We consider the model with wide variety of subject specific

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, \cdots, N, \tag{1.1}$$

where

- $\boldsymbol{y}_i$ is an $n_i \times 1$ vector of responses of the ith subject; also called individual or cluster.
- $\boldsymbol{X}_i = \boldsymbol{Z}_i\boldsymbol{K}_i$ is an $n_i \times p$ design matrix of explanatory variables; also called covariates or fixed effects.
- $\boldsymbol{\beta}$ is an $p \times 1$ vector of population parameters; also called covariates or fixed effects coefficients.
- $\boldsymbol{Z}_i$ is an $n_i \times q$ design matrix of random effects.
- $\boldsymbol{\varepsilon}_i$ is an $n_i \times 1$ error term with independent components, each of them has zero mean and the within subject variance σ^2.
- $\boldsymbol{b}_i$ is an $q \times 1$ vector of random effects with zero mean and covariance matrix $\boldsymbol{D}_*$.

It assumed that all random vectors $\{\boldsymbol{b}_i, \boldsymbol{\varepsilon}_i, i = 1, \cdots, N\}$ are mutually independent. To make the LME model identifiable, which provides the uniqueness of the distribution as a function of parameters, we assume that matrix $\sum \boldsymbol{X}_i^T \boldsymbol{X}_i$ is nonsingular; $\sum n_i > p$; at least one matrix $\boldsymbol{Z}_i^T \boldsymbol{Z}_i$ is positive definite and $\sum_{i=1}^{N} (n_i - q) > 0$ (Demidenko (2004)).

The model contains fixed effects coefficients $\boldsymbol{\beta}$ constant for all subjects and random effects $\boldsymbol{b}_i$ specific for the ith subject. Because the model comprises fixed and random effects, it is termed a mixed effects model. While the population average model assumes that the marginal variance of the response variable is homogeneous in the population, subject specific model allows the marginal variance of the response variable to be heterogenous. However, we usually make a homogeneity assumption for the within-subject error $\boldsymbol{\varepsilon}_i$, or the conditional variance of the response variable given the random effects. For simplicity, people usually assume that $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_i$ are normally distributed as

$$\boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \qquad \boldsymbol{b}_i \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{D}),$$

where $\boldsymbol{I} = \boldsymbol{I}_{n_i}$ is identity matrix of dimension n_i. And the linear mixed effects model under this assumption can be written in marginal form as

$$\boldsymbol{y}_i \sim N(\boldsymbol{X}_i \boldsymbol{\beta}, \sigma^2 (\boldsymbol{I} + \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T)), \qquad i = 1, \cdots, N.$$

The standard residual plots can be used to assess this assumption by computing

$$\hat{\epsilon}_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \mathbf{X}_{ij}\hat{\boldsymbol{\beta}} - \mathbf{Z}_{ij}\hat{\mathbf{b}}_i,$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}_i$ are the estimated fixed and random effects. The plot of $\hat{\epsilon}_{ij}$ versus $\hat{Y}_{ij}$ should show no pattern if the assumption holds. Many textbooks introduce this procedure for model diagnosis (e.g. Pinheiro and Bates (2000), Demidenko (2004)).

Residual analysis is usually conducted to check assumptions for linear regression models. Checking model assumption in mixed-effects model has been considered by many authors in recent decades. Most previous works are concerned with model mis-specification for mean functions, model robustness against distributional assumption and detection of influential observations (Demidenko (2004), Albert (2008), Neuhaus et al. (2011), Benjamin and Amy (2009), among others). However, there appears to be not much discussion or systematic procedure for testing a common covariance matrix for the subject effects in a mixed-effects models. In model (1.1), the between subject variation has also been assumed to follow a common covariance structure for all subjects and there is not much in the literature about ascertaining whether such an assumption is plausible or valid. As far as we know, there seems to be no statistical procedure that can be directly used to detect violation of this assumption. The equal covariance assumption for random

effects is actually as strong or arguably stronger than the equal variance assumption commonly made for the simple linear regression model. When the assumption is not met in a data analysis, all kinds of conventional hypothesis tests and inference procedures may be misleading. Our aim here is to propose tests for checking this homogeneity assumption in a mixed-effects model and ascertain their performance. As we demonstrate with numerical results in this thesis, the magnitude and significance of the regression coefficients may be affected in the presence of outliers.

This idea also arises from a substantial interest. The rheumatoid arthritis disease is an important research topic for medical literature for a long time. We consider a cohort study which investigates the impacts of rheumatoid arthritis on health and life quality (as measured by health assessment questionnaire score in Kirwan and Reeback (1986)) for the individuals. The scientific interest is to identify risk factors that predict the disease outcomes in patients with rheumatoid arthritis. Many previous studies only used cross-sectional analysis (e.g. Gordon et al. (2001)) or focused on the outcome's change between baseline and endpoint (e.g. Tanaka et. al. (2008)). The outcome variables in our example are measured repeatedly over several years follow-up for each subject, and the random effect model is a standard method for incorporating subject-to-subject variation in the longitudinal data. Furthermore, almost all earlier studies analyzed either the

continuous outcome or the count outcome while we attempt to jointly model the two outcomes in a single model by assuming a correlated structure for the random effects. Consequently, it is rather imperative to ensure the assumption for the variance of random effects across the study population in our sophisticated model.

In this thesis, we provide formal methodology to check homogeneity assumption in mixed-effects models, following a review of generalized linear mixed-effects models. We also conduct simulation studies to examine the performance of the proposed test. In addition, a medical example of arthritis disease study is analyzed to illustrate our methods.

CHAPTER 2

Testing of Homogeneity Hypothesis in Mixed Models

2.1 The Linear Mixed effects model

2.1.1 The Log-likelihood functions

Recall the model (1.1), under the assumptions in chapter 1, The LME model with normally distributed random variables can be written in marginal form as

$$\boldsymbol{y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2(\boldsymbol{I} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T)), \qquad i = 1, \cdots, N. \tag{2.1}$$

By the marginal form of the LME model, after dropping the constant term $C = -(N_T/2)\ln(2\pi)$, the log-likelihood function is given by

$$l(\boldsymbol{\theta}) = -\frac{1}{2}\{N_T \ln \sigma^2 + \sum_{i=1}^{N} [\ln |\boldsymbol{I} + \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T| + \sigma^{-2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T (\boldsymbol{I} + \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T)^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})]\}$$

where $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^T, \sigma^2, vech^T(\boldsymbol{D})\right)^T$ is a combined vector of known parameters and N_T is the total number of all the observations. And $vech(\boldsymbol{D})$ denotes the $q(q+1)/2$ vector of unique elements of symmetric matrix $\boldsymbol{D}$ (Magnus (1988)). Therefore, the total dimension of the parameter vector $\boldsymbol{\theta}$ is $p + 1 + q(q+1)/2$ and the Maximum Likelihood Estimate (MLE) maximizes function l over the parameter space

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} : \boldsymbol{\beta} \in R^p, \sigma^2 > 0, \boldsymbol{D} \text{ is nonnegative definite}\}.$$

The scaled covariance matrix for $\boldsymbol{y}_i$ is given by

$$\boldsymbol{V}_i = \boldsymbol{I} + \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T$$

By using the notation $\boldsymbol{V}_i$, the log likelihood function can be written as

$$l(\boldsymbol{\theta}) = -\frac{1}{2}\{N_T \ln \sigma^2 + \sum_{i=1}^{N} [\ln |\boldsymbol{V}_i| + \sigma^{-2} \boldsymbol{e}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{e}_i]\},$$

where

$$\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}$$

is an $n_i \times 1$ residual vector for the ith cluster, $i = 1, \cdots N$.

To estimate the parameters, we need to maximize the log likelihood function. Using standard formulas we obtain

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sigma^{-2} \sum \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}),$$

$$\begin{aligned}\frac{\partial l}{\partial \sigma^2} &= -\frac{1}{2}N_T\sigma^{-2} + \frac{1}{2}\sigma^{-4}\sum(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}),\\ \frac{\partial l}{\partial \boldsymbol{D}} &= -\frac{1}{2}\sum[\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i - \sigma^{-2}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i].\end{aligned}$$

The maximum likelihood estimate solves the system of equations

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{0}, \quad \frac{\partial l}{\partial \sigma^2} = 0, \quad \frac{\partial l}{\partial \boldsymbol{D}} = \mathbf{0},$$

for $\boldsymbol{\beta}, \sigma^2$ and $\boldsymbol{D}$ (Demidenko (2004)).

There are three general types of algorithms in statistics to maximize the log-likelihood function: Newton-Raphson (NR), Fisher scoring (FS) and Expectation-Maximization (EM). The NR algorithm may fail if the starting point is far from the maximum, but it is fast when the starting point is relatively close to the maximum. The FS algorithm is more robust to the choice of starting point since the information matrix is always positive definite. And the EM algorithm may be slow when the matrix $\boldsymbol{D}$ close to zero.

Pinheiro (1994) has shown that, under certain regularity conditions generally satisfied in practice, the maximum likelihood estimates in the LME model are consistent and asymptotically normal where the approximate variance-covariance matrix for the maximum likelihood estimates is given by the inverse of the information matrix (Cox and Hinkley (1974)).

The parameters in the LME model are $\boldsymbol{\beta}, \sigma^2, vech(\boldsymbol{D})$, the information matrix

for the parameters $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^T, \sigma^2, vech^T(\boldsymbol{D})\right)^T$ is given by

$$\boldsymbol{I} = \begin{bmatrix} \sigma^2 \sum \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0.5 N_T \sigma^{-4} & 0.5\sigma^{-2} \sum vec^T(\boldsymbol{R}_i) \\ \mathbf{0} & 0.5\sigma^{-2} \sum vec(\boldsymbol{R}_i) & 0.5 \sum \boldsymbol{R}_i \otimes \boldsymbol{R}_i \end{bmatrix},$$

(Demidenko (2004)) where the $q \times q$ matrix $\boldsymbol{R}_i$ is defined as

$$\boldsymbol{R}_i = \boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{Z}_i = \boldsymbol{Z}_i^T (\boldsymbol{I} + \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T)^{-1} \boldsymbol{Z}_i.$$

2.1.2 Estimation and Inference

Known variance

When all covariance parameters in the marginal distribution are known, the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ is given by

$$\tilde{\boldsymbol{\beta}} = (\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{y}_i \tag{2.2}$$

where $\boldsymbol{W}_i$ equals $\sigma^{-2}\boldsymbol{V}_i^{-1}$. The estimate of $\boldsymbol{b}_i$ is empirical Bayes with the form

$$\tilde{\boldsymbol{b}_i} = E[\boldsymbol{b}_i | \boldsymbol{Y}_i = \boldsymbol{y}_i] = \int \boldsymbol{b}_i f(\boldsymbol{b}_i | \boldsymbol{y}_i) d\boldsymbol{b}_i = \boldsymbol{D} \boldsymbol{Z}_i^T \boldsymbol{W}_i (\boldsymbol{y}_i - \boldsymbol{X}_i \tilde{\boldsymbol{\beta}}) \tag{2.3}$$

when $\boldsymbol{y}_i$ is given. And this is also the Best Linear Unbiased Prediction (BLUP) (Robinson (1991)).

Since both $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{b}}$ are linear functions of $\boldsymbol{y}_i$, and $\boldsymbol{y}_i$ has marginal distribution

(2.1), the expressions for their variance can be easily derived as

$$var(\tilde{\boldsymbol{\beta}}) = (\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i)^{-1} \tag{2.4}$$

and

$$var(\tilde{\boldsymbol{b}_i}) = \boldsymbol{D}\boldsymbol{Z}_i^T \{\boldsymbol{W}_i - \boldsymbol{W}_i \boldsymbol{X}_i (\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i)^{-1} \boldsymbol{X}_i^T \boldsymbol{W}_i\} \boldsymbol{Z}_i \boldsymbol{D}. \tag{2.5}$$

If $var(\tilde{\boldsymbol{b}_i})$ is used to assess the error of estimation, the variability in $\tilde{\boldsymbol{b}_i} - \boldsymbol{b}_i$ is underestimated since it ignores the variation of $\boldsymbol{b}_i$. Therefore, the inference for $\boldsymbol{b}_i$ is sometimes based on

$$var(\tilde{\boldsymbol{b}_i} - \boldsymbol{b}_i) = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{Z}_i^T \boldsymbol{W}_i \boldsymbol{Z}_i \boldsymbol{D} + \boldsymbol{D}\boldsymbol{Z}_i^T \boldsymbol{W}_i \boldsymbol{X}_i (\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i)^{-1} \boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{Z}_i \boldsymbol{D}.$$

Unknown variance

When the covariance matrix is unknown, but an estimate of the parameters in covariance matrix is available, it is natural to set

$$\hat{\sigma}^2 \hat{\boldsymbol{V}}_i = \hat{\sigma}^2 (\boldsymbol{I} + \boldsymbol{Z}_i \hat{\boldsymbol{D}} \boldsymbol{Z}_i^T) = \hat{\boldsymbol{W}}_i^{-1}$$

and estimate $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ by

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{N} \boldsymbol{X}_i^T \hat{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i^T \hat{\boldsymbol{W}}_i \boldsymbol{y}_i$$

and

$$\hat{\boldsymbol{b}_i} = \hat{\boldsymbol{D}} \boldsymbol{Z}_i^T \hat{\boldsymbol{W}}_i (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}})$$

substituting all the unknown variance-covariance parameters by their maximum likelihood estimates into equations (2.2) and (2.3). We can also apply the same

idea substituting unknown parameters into equation (2.4) and (2.5) to get $var(\hat{\boldsymbol{\beta}})$ and $var(\hat{\boldsymbol{b}_i})$, which is given by

$$var(\hat{\boldsymbol{\beta}}) = (\sum_{i=1}^{N} \boldsymbol{X}_i^T \hat{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1}$$

and

$$var(\hat{\boldsymbol{b}_i}) = \hat{\boldsymbol{D}} \boldsymbol{Z}_i^T \{\hat{\boldsymbol{W}}_i - \hat{\boldsymbol{W}}_i \boldsymbol{X}_i (\sum_{i=1}^{N} \boldsymbol{X}_i^T \hat{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1} \boldsymbol{X}_i^T \hat{\boldsymbol{W}}_i\} \boldsymbol{Z}_i \hat{\boldsymbol{D}}.$$

This approach is maximum likelihood for $\boldsymbol{\beta}$ and empirical Bayes for $\boldsymbol{b}_i$ when we consider the estimation of all the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2, vech^T(\boldsymbol{D}))^T$ simultaneously by maximizing their joint likelihood function based on the marginal distribution of $\boldsymbol{y}_i$.

2.2 Generalized Linear Mixed Models

2.2.1 Introduction

LME models have been widely used in situations where the observations are continuous. However, there are many cases in practice where the observations are discrete or categorical. Nelder and Wedderburn (1972) proposed an extension of linear models, called generalized linear model, or GLM. In the classical linear

models, the mean of the observation is a linear function of some covariates and the variance of the observation is a constant. In the extension to GLM, some modification should be done to these conditions. In contrast, the mean of the observation is associated with a linear function of some covariates through a link function and the variance of the observation is a function of the mean in GLM. Unlike linear models, GLMs include a variety of models that includes normal, binomial, Poisson and multinomial as special cases. And overdispersion which is the presence of greater variability in a data set than would be expected based on a given statistical model is relatively common in real life regression problem with Poisson and multinomial models.

Generalized linear mixed-effects models (GLMM) combine the ideas of generalized linear models with the random effects modeling ideas. The response random variables Y_{ij} $(i = 1, \cdots, N; j = 1, \cdots, n_i)$ are conditional independent given the random effects $\mathbf{b}_i$, each following an exponential family distribution

$$f(y_{ij}|\theta_{ij}, \phi) = \exp\left[\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a(\phi)} + c(y_{ij}, \phi)\right]$$

where $b(\cdot), a(\cdot), c(\cdot, \cdot)$ are known functions and ϕ is a dispersion parameter (Berridge and Crouchley (2011)). The canonical parameter θ_{ij} is associated with the conditional mean $\mu_{ij} = E(y_{ij}|\mathbf{b}_i)$ where $\mathbf{b}_i$ are random effects. The conditional mean further depends upon covariates via the linear predictor η_{ij} using the link function

g by $\eta_{ij} = g(\mu_{ij})$ where

$$\eta_{ij} = \mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\mathbf{b}_i.$$

We assume that $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}_*)$ and note that under the so-called canonical link $\theta_{ij} = \eta_{ij}$. GLMM can be used to model response variables with normal, binomial and Poisson distributions.

It is straight forward to construct the likelihood and estimate model parameters by maximizing the likelihood. However, the likelihood function under a GLMM usually does not have a closed form solution. Although MLE and RMLE methods are standard procedures for a normal response, likelihood based inference for a non-normal response is computationally challenging. Many advanced approaches have been developed to solve the computational difficulties, which includes Monte Carlo EM algorithm (e.g. McCulloch (1997); Booth and Hobert (1999)), nonlikelihood based computationally attractive methods (e.g. Breslow and Clayton (1993)), generalized estimating equation (GEE; e.g. Diggle et al. (1996)) and Bayesian method based on the Gibbs sampler (e.g. Zeger and Karim (1991)).

2.2.2 Overdispersion

Binary data

In statistics, dispersion denotes how stretched or squeezed a distribution is and

overdispersion is the presence of greater variability in a data set than would be expected based on a given simple statistical model. In case of binary data models, binomial distribution is widely used in which the observations $Y_{ij} \sim bin(n_{ij}, p_{ij})$ and the variance is $Var(y_{ij}) = n_{ij}p_{ij}(1 - p_{ij})$ given the number of experiments n_{ij} and success probability p_{ij}. However, in some real world application cases, when overdispersion is present the variance will be greater than the one mentioned above. There are generally three main approaches modeling the overdispersion.

(a) Multiplicative overdispersion models

In a multiplicative overdispersion model, which is also called scale or constant overdispersion model, the variance of the binomial distribution is modified by $Var(y_{ij}) = \phi n_{ij}p_{ij}(1 - p_{ij})$, where ϕ is a constant dispersion parameter. If we call the modified binomial model as overdispersed binomial, the simplest overdispersed binomial GLMM with logit link can be written as:

$$
\begin{aligned}
Y_{ij} &\sim overdispersed\ binomial(n_{ij}, p_{ij}, \phi) \\
p_{ij} &= logit^{-1}(\beta + b_i) \\
b_i &\sim N(0, \sigma^2)
\end{aligned}
$$

where ϕ is the dispersion parameter and σ^2 is the variance of random effects. In this standard quasi-likelihood approach, the estimates for the linear predictor parameters are the same as those for the simple binomial GLMM, while the overdispersion factor ϕ is estimated by dividing the Pearson statistic χ^2 by its degrees of freedom

(McCullagh and Nelder (1989)).

(b) Additive overdispersion models

As an alternative to the multiplicative model, the overdispersion can be modelled as a residual term on the link scale (Browne et al. (2005)). In this case, the simplest binomial GLMM can be written as

$$Y_{ij} \sim binomial(n_{ij}, p_{ij})$$

$$p_{ij} = logit^{-1}(\beta + b_i + e_{ij})$$

$$b_i \sim N(0, \sigma^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

where e_{ij} is the additive overdispersion term on the link scale with variance σ_e^2. In such models, underdispersion is not allowed if the variance σ_e^2 is constrained to be non-negative.

(c) Beta-binomial variance function model

There is another more complicated two stage model called beta-binomial variance function model, in which we assume that $Y_{ij} \sim bin(n_{ij}, P_{ij})$ where the P_{ij}'s are taken as random variables with $E(P_{ij}) = \pi_{ij}$ and $Var(P_{ij}) = \phi\pi_{ij}(1 - \pi_{ij})$. A special case of this is the beta binomial distribution by assuming $P_{ij} \sim beta(a_{ij}, b_{ij})$ with $\frac{a_{ij}}{a_{ij}+b_{ij}} = \pi_{ij}$. In such models, quasi-likelihood equations can be defined and the maximum quasi-likelihood estimates can be obtained by iterated reweighted least squares.

Count data

For count data which are extremely common in real application, various types of Poisson mixed models have been proposed. We now assume that the random variables Y_{ij} represents counts with mean μ_{ij}. The standard Poisson model assumes that $Y_{ij} \sim Pois(\mu_{ij})$ with variance function $var(Y_{ij}) = \mu_{ij}$. As for the binary data, when overdispersion occurs, there are generally three approaches.

(a) Multiplicative overdispersion models

When there is overdispersion we need to consider variance functions which predict greater variability. A simple constant overdispersion model replaces the Poisson variance by $var(Y_{ij}) = \phi\mu_{ij}$. If we call the modified Poisson model as overdispersed Poisson, then the simplest Possion GLMM with multiplicative overdispersion can be written as:

$$Y_{ij} \sim overdispersed\ Poisson(\mu_{ij}, \phi)$$

$$\mu_{ij} = exp(\beta + b_i)$$

$$b_i \sim N(0, \sigma^2)$$

where ϕ is the dispersion parameter and σ^2 is the variance of random effects. The estimation procedure is the same as for the logit binary model.

(b) Additive overdispersion models

As an alternative to the multiplicative model, the overdispersion can be modelled as a residual term on the link scale (Browne et al. (2005)). In this case, the

simplest binomial GLMM can be written as

$$Y_{ij} \sim Poisson(\mu_{ij})$$

$$p_{ij} = logit^{-1}(\beta + b_i + e_{ij})$$

$$b_i \sim N(0, \sigma^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

where e_{ij} is the additive overdispersion term on the link scale with variance σ_e^2. In such models, underdispersion is not allowed if the variance σ_e^2 is constrained to be non-negative.

(c) Negative binomial type variance model

A two stage model as in binary case is called negative binomial type variance model, in which we assume that $Y_{ij} \sim Pois(\theta_{ij})$ and θ_{ij} are random variables with $E(\theta_{ij}) = \mu_{ij}$ and $Var(\theta_{ij}) = \sigma_{ij}^2$. A simple case is that θ_{ij} follows a $Gamma(k, \lambda_{ij})$ which leads to a negative binomial distribution for Y_{ij} with $E(Y_{ij}) = k\lambda_{ij} = \mu_{ij}$ and $Var(Y_{ij}) = \mu_{ij} + k\mu_{ij}^2$. The maximum likelihood estimation can be obtained by the iteratively reweighted least square algorithm for generalized linear models.

2.2.3 Variance Partition Coefficients

The accommodation of random effects with GLMMs suggests the calculation of variance partition coefficients which represents group or macro level variation in

an outcome variable as a proportion of total variation and allows for conditioning on covariate values (Li et al. (2008)). The VPC parameter is linked to the widely used intra-class correlation coefficient (ICC), a measure that typically indicates the proportion of among-group variation in intercept-only linear models and the estimation of VPCs is usually based on fitting sophisticatedly structured models (Snijders and Bosker (1999)).

We consider a three level logistic regression model for hierarchical binary response data. Let Y_{ijk} be a Bernoulli random variable for the kth observation on the jth individual within the ith cluster,$i = 1, \ldots, m; j = 1, \ldots, n_i; k = 1, 2, \ldots, n_{ij}$. Here, $n = \sum_{i=1}^{m} n_i$ represents the total number of individuals, and $N = \sum_i \sum_j n_{ij}$ the total number of observations. The model is given by

$$\log \frac{p_{ijk}}{1 - p_{ijk}} = \eta_{ijk} = \boldsymbol{x}_{ijk}^T \beta + \boldsymbol{z}_i^T \boldsymbol{u}_i + \boldsymbol{z}_{ij}^T \boldsymbol{v}_{ij}$$

where $\boldsymbol{z}_i = (1, z_{i1}, \ldots, z_{ip})^T$ and $\boldsymbol{z}_{ij} = (1, z_{ij1}, \ldots, z_{ijp})^T$ are covariates for the random effects. The random coefficients $\boldsymbol{u}_i$ and $\boldsymbol{v}_{ij}$ are conditionally independent and with multivariate normal distributions, $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{\Omega_u}), \boldsymbol{v}_{ij} \sim N(\boldsymbol{0}, \boldsymbol{\Omega_v})$. All random components are assumed independent.

Four methods of VPCs for nonlinear two-level models have been proposed previously: Taylor series linearization method, simulation-based method, latent variable

method, and the naive linear model method (Goldstein (2003)). As each definition arises from different assumptions, each definition may lead to different VPC values.

(a) *Linearization.* The first definition of VPC arises when we approximate the binary outcome variable with linear representations under the assumed model. The unconditional VPCs can be defined as

$$E[\tau_u(\boldsymbol{x}_{ijk}, \boldsymbol{x}_{ij'k'})] = \int \tau_u(\boldsymbol{x}_{ijk}, \boldsymbol{x}_{ij'k'}) dF(\boldsymbol{x}_{ijk}, \boldsymbol{x}_{ij'k'})$$

and

$$E[\tau_v(\boldsymbol{x}_{ijk}, \boldsymbol{x}_{ij'k'})] = \int \tau_v(\boldsymbol{x}_{ijk}, \boldsymbol{x}_{ij'k'}) dF(\boldsymbol{x}_{ijk}, \boldsymbol{x}_{ij'k'})$$

where $F(\cdot, \cdot)$ is the joint probability distribution of two covariates, τ_u is the VPC at level 3 and τ_v is the VPC at level 2 . The practical choice of F can be based on an empirical distribution estimator or a prior information.

(b) *Simulation-Based.* The second VPC definition is obtained by re-constructing the data generation process from computer simulations and recording the observed variation. This method is conceptually natural and computationally stable, with accuracy being an increasing function of the number of simulations.

(c) *Latent-Response Model.* Logistic regression model has a close relationship with the latent variable model. We may write the model as a latent response

model

$$y^*_{ijk} = \boldsymbol{x}^T_{ijk}\beta + u_{i0} + \sum_{l=1}^{p} z_{il}u_{il} + v_{ij0} + \sum_{l=1}^{q} z_{ijl}v_{ijl} + \epsilon_{ijk}$$

where ϵ_{ijk} has a standard logistic distribution with variance $\pi^2/3$ and the observed responses are generated from the threshold model

$$y_{ijk} = \begin{cases} 1 & if \;\; y^*_{ijk} > 0 \\ 0 & otherwise. \end{cases}$$

For the same cluster i, but different individuals j and $j^{'}$, define level 3 VPC as

$$\tau_u = L_{1i}/[(L_{1i} + L_{2ij} + \pi^2/3)(L_{1i} + L_{2ij'} + \pi^2/3)]^{1/2}$$

whereas, for the same individual j in cluster i but different observations, define level 2 VPC as

$$\tau_v = (L_{1i} + L_{2ij})/(L_{1i} + L_{2ij} + \pi^2/3)$$

where $L_{1i} = \boldsymbol{z}_i^T\boldsymbol{\Omega}_u\boldsymbol{z}_i$ is the variance at level 3 while $L_{2ij} = \boldsymbol{z}_{ij}^T\boldsymbol{\Omega}_v\boldsymbol{z}_{ij}$ is the conditional variance at level 2.

If all $\boldsymbol{z}$ for random effects only include random intercepts, the resulting τ_u and τ_v remain constant across all the samples.

(d) *Binary Linear Model.* As an approximation, we may treat the $0, 1$ responses as if they are normally distributed variables and estimate the VPC under a full linearity assumption. The model is written as

$$y_{ijk} = \boldsymbol{x}^T_{ijk}\beta^* + \boldsymbol{z}_i^T\boldsymbol{u}^*_i + \boldsymbol{z}_{ij}^T\boldsymbol{v}^*_{ij} + \epsilon^*_{ijk}$$

where $\boldsymbol{u}_i^* \sim N(\mathbf{0}, \boldsymbol{\Omega}_u^*), \boldsymbol{v}_{ij}^* \sim N(\mathbf{0}, \boldsymbol{\Omega}_v^*), \epsilon_{ijk}^* \sim N(0, \sigma_\epsilon^{*2})$ The VPCs at the two levels are then

$$\tau_u = L_{1i}^*/[(L_{1i}^* + L_{2ij}^* + \sigma_\epsilon^{*2})(L_{1i}^* + L_{2ij'}^* + \sigma_\epsilon^{*2})]^{1/2},$$

$$\tau_v = (L_{1i}^* + L_{2ij}^*)/(L_{1i}^* + L_{2ij}^* + \sigma_\epsilon^{*2})$$

where $L_{1i}^* = \boldsymbol{z}_i^T \boldsymbol{\Omega}_u^* \boldsymbol{z}_i, L_{2ij}^* = \boldsymbol{z}_{ij}^T \boldsymbol{\Omega}_v^* \boldsymbol{z}_{ij}$. This definition is easy to implement since we only need to fit a linear instead of nonlinear mixed effects models. But it fails when the true response probability close to 0 or 1.

Selection of a VPC calculation method should refer to a data generation process. The simulation-based approach (Definition 2) acknowledges random variation in data generation process and thus should be recommended for a scrupulous variance study. We remark that Definition 1 is built upon the idea of approximating the marginal variance of the binary response and thus has a theoretical advantage over other methods. Definition 3 is attractive because of its simplicity as long as steady performance. Lastly, the linear approximation method (Definition 4) may be used to provide approximations when the marginal response probability $P(Y = 1)$ is not extreme.

2.3 R function

For linear mixed models, function *lme* from package `nlme` is widely used for its convenience and stability, but it is not applicable for GLMMs. Function *glmer* from package `lme4` is a good choice when we fit GLMM models. It handles crossed random effects as well and does not require complex variance structure. Overdispersion arises in many real world regression problems. Other than *glmer* function which uses Laplace approximation and adaptive Gauss-Hermite quadrature method, the function $glmmPQL$ from package `MASS` implements multiplicative dispersion GLMM fitted with penalized quasi-likelihood estimation while the function $MCMCglmm$ from package `MCMCglmm` implements additive overdispersion GLMM fitted by MCMC sampling from the posterior distribution. The restricted maximum likelihood estimation (RMLE) based procedures use approximate likelihood method may not work well for GLMM since analytical results for non-Gaussian GLMM are generally not available. The accuracy of MCMC procedure increases the longer the analysis is run for although it can be slow and technically challenging.

2.4 Test for Homogeneity

2.4.1 Testing of the presence of random effects

Deciding which random effects may vary across subjects is an important issue. It is often of interest to test whether certain random effects should be included, which means to test whether the variance of random effects equal to 0. In statistical language, this translates into hypothesis testing,

$$H_0 : \boldsymbol{D} = \boldsymbol{0}.$$

Since $\boldsymbol{D} = \boldsymbol{0}$ is the boundary point of the parameter space, one can expect that the actual significance level of the likelihood ratio test will be less than nominal. An exact F-test was developed for the hypothesis for the LME model. The idea of the test is that when $\boldsymbol{D} = \boldsymbol{0}$, the difference between the minimum sum of squares with random effects, S_{min} and the minimum sum of squares without random effects(OLS) should be close. Thus, we computer the residual sum of squares

$$S_{OLS} = \sum_{i=1}^{N} \left| \boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}_{OLS} \right|^2$$

assuming no random effects. Let $r = rank(\boldsymbol{W})$, $N_T = \sum_{i=1}^{N} n_i$, m be de dimensions of fixed effects, the ratio of the two quadratic forms has an F-distribution, or more

precisely

$$\frac{(S_{OLS} - S_{min})/(r-m)}{S_{min}/(N_T - r)} \sim F(r-m, N_T - r)$$

When random effects are present in the LME model, S_{min} should be relatively small, and therefore the ratio becomes large. Thus, we reject $H_0 : \boldsymbol{D} = \mathbf{0}$ if the ratio is large. More precisely, let $1-\alpha$ be the chosen significance level, e.g., $\alpha = 0.05$, and $f_{0.95}$ be the 0.95 quantile of F-distribution with $r-m$ and $N_T - r$ degrees of freedom, then H_0 is rejected when

$$\frac{(S_{OLS} - S_{min})/(r-m)}{S_{min}/(N_T - r)} > f_{0.95}$$

The classical procedures such as the likelihood ratio test for a single variance component (e.g. Self and Liang (1987); Stram and Lee (1994)) can be carried out using mixtures of chi-square distribution. For multivariate tests, distribution of the test statistics are not simple. Many alternative frequentist methods are studied including score tests (e.g. Commenges and Jacqmin-Gadda (1997), Verbeke and Molenberghs (2003)), Wald tests (e.g. Silvapulle (1992), Molenberghs and Verbeke (2007)) and generalized likelihood ratio tests (e.g. Crainiceanu and Ruppert (2004)). Another approach for testing random effects using approximate Bayes factors is known to encounter the difficulty of multivariate tests (Benjamin and Amy (2009)).

2.4.2 Membership testing

The following membership problem is sometimes of interest: let $y_1, y_2, \ldots, y_n$ be an independent and identically distributed sample from a general population and y_{n+1} be a new observation, independent of the previous n observations. Does y_{n+1} belong to the same population. This is a typical question in medical diagnostics, in which case the y_i are the observations of normal patients and y_{n+1} is the observation of a new patient.

Recall that for the linear hypothesis in a general linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta}, \boldsymbol{\eta} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{V}),$$

where σ^2 is unknown but the covariance matrix $\boldsymbol{V}$ is known and nonsingular, and $\boldsymbol{X}$ is the $n \times m$ design matrix of full rank. We want to test the linear hypothesis $H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0}$ where $\boldsymbol{C}$ is a fixed $q \times m$ matrix of full rank. Define two residual sums of squares,

$$RSS = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), RSS_0 = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{0}})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{0}})$$

where RSS is the absolute minimum of the weighted least squares and $\hat{\beta}$ is the GLS estimate. RSS_0 is the residual sum of the weighted least squares under restriction $\boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0}$ and $\hat{\beta}_0$ is the GLS estimate under restriction. Then under H_0,

$$\frac{(RSS_0 - RSS)/q}{RSS/(n-m)} \sim F(q, n-m),$$

(Rao and Toutenburg (1999)). In the LME membership test, we have N normal patients and a $(N+1)$th patient who follows the same model but possibly with different fixed effect coefficients,

$$\boldsymbol{y}_{N+1} = \boldsymbol{X}_{N+1}\boldsymbol{\beta}_* + \boldsymbol{Z}_{N+1}\boldsymbol{b}_{N+1} + \epsilon_{N+1}$$

where $\boldsymbol{b}_{N+1} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{D})$ and $\epsilon \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. And the membership problem can be translated into the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_*$. In this case, the minimal weighted sum of squares under H_0 is

$$RSS_0 = \sum_{i=1}^{N+1} (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_{N+1})' \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_{N+1})$$

and the minimal weighted sum of squares is

$$RSS = \sum_{i=1}^{N} (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})' \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) + (\boldsymbol{y}_{N+1} - \boldsymbol{X}_{N+1}\hat{\boldsymbol{\beta}}_*)' \boldsymbol{V}_{N+1}^{-1} (\boldsymbol{y}_{N+1} - \boldsymbol{X}_{N+1}\hat{\boldsymbol{\beta}}_*)$$

where

$$\hat{\boldsymbol{\beta}}_* = (\boldsymbol{X}_{N+1}' \boldsymbol{V}_{N+1}^{-1} \boldsymbol{X}_{N+1})^{-1} \boldsymbol{X}_{N+1}' \boldsymbol{V}_{N+1}^{-1} \boldsymbol{y}_{N+1}.$$

Hence, according to the F-test, we say that the new patient is normal if

$$\frac{(RSS_0 - RSS)/m}{RSS/(\sum n_i - m)} \sim F(m, \sum n_i - m).$$

2.4.3 Testing of Homogeneity Hypothesis

We now consider a formal test for the homogeneity assumption in GLMM. That is, we test the assumption that all the random effects following the same distribution. If we can find some subjects or clusters with abnormal random effects which contradict the model assumption, we may conclude that such subjects/clusters are different from others in the population and can be treated as outliers. The hypotheses are stated as

$$H_0 : var(\boldsymbol{b}_i) = \boldsymbol{D} \quad vs \quad H_1 : var(\boldsymbol{b}_i) \neq \boldsymbol{D}.$$

The test is based on the model-calibrated observations $\tilde{\mathbf{b}}_i$. To examine the distribution of $\mathbf{b}_i$, we are lack of actual observations and have to rely on the model prediction. This approach is not entirely new for mixed-effects research and appeared before for other purposed (Pinheiro and Bates (2000)).

Recall section 2.1.2, for normal response, it is well-known that $\tilde{\mathbf{b}}_i$ is the best linear unbiased prediction (BLUP) (Robinson (1991)) and is given by

$$\tilde{\boldsymbol{b}_i} = E[\boldsymbol{b}_i|\boldsymbol{Y}_i] = \boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{W}_i(\boldsymbol{Y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}})$$

where $\boldsymbol{W}_i = \sigma^{-2}(\boldsymbol{I} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T)^{-1}$. It is also known that

$$\Sigma = \text{var}(\tilde{\boldsymbol{b}_i}) = \boldsymbol{D}\boldsymbol{Z}_i^T\{\boldsymbol{W}_i - \boldsymbol{W}_i\boldsymbol{X}_i(\sum_{i=1}^{N}\boldsymbol{X}_i^T\boldsymbol{W}_i\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^T\boldsymbol{W}_i\}\boldsymbol{Z}_i\boldsymbol{D}.$$

We denote var($\tilde{\boldsymbol{b}_i}$) as Σ. When a model is fitted under the homogeneity assumption, we obtain an estimate $\hat{\mathbf{D}}$ and can then replace $\mathbf{D}$ in (2.3) and (2.5) with its estimate. The resulting estimates are denoted by $\hat{\mathbf{b}}_i$ and $\hat{\Sigma}$ in the thesis. For non-normal response, the exact form of $\tilde{\mathbf{b}}_i$ is complicated and is no longer a linear function of Y since the posterior of $\mathbf{b}_i$ given $\mathbf{Y}_i$ is complicated. Nonetheless, $\hat{\mathbf{b}}_i$ and $\hat{\Sigma}$ for GLMM may still be easily calculated with a numerical method from many statistical packages.

Recall that if $\boldsymbol{Y}$ is a k-dimensional Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}$, then $\boldsymbol{X} = (\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})$ follows a chi-square distribution with k degree of freedom. A Wald-type test may be constructed as

$$T_i = \hat{\boldsymbol{b}_i}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{b}_i} \tag{2.6}$$

which asymptotically follows a chi-square distribution with q degrees of freedom under the null hypothesis. We reject the null hypothesis that the random effect for cluster i has the assumed covariance matrix when the test statistic is large. The result of a small p-value may be caused by abnormal cluster or misspecifying the random effects following normal distribution. The covariance matrix estimate $\hat{\Sigma}$ is obtained from maximizing the assumed model likelihood. Therefore we call (2.6) model based test.

On the other hand, a common empirical estimator for Σ may be constructed by

$$\hat{\boldsymbol{\Sigma}^*} = \frac{1}{N} \sum_{i=1}^{N} \hat{\boldsymbol{b}_i} \hat{\boldsymbol{b}_i}^T$$

which is just the sample variance of $\hat{\boldsymbol{b}}_i$ under the model assumption. We then propose another test

$$T_i^* = \hat{\boldsymbol{b}}_i^{\,T} \hat{\boldsymbol{\Sigma}}^{*\,-1} \hat{\boldsymbol{b}}_i \tag{2.7}$$

which asymptotically also follows a chi-square distribution with degrees of freedom q under the null. We refer to (2.7) as an empirical test in this thesis. A simulation study is carried out for evaluating the performance of the model based test and empirical test.

2.5 Simulation

2.5.1 Normal response with a random intercept

We simulated 500 data sets based on a linear mixed effects model with a random intercept

$$y_{ij} = \beta_1 + X_{ij1}\beta_2 + X_{ij2}\beta_3 + b_i + \varepsilon_{ij}, \qquad i = 1, \cdots, N; j = 1, \cdots, n_i$$

The β's are set to be fixed values. We generate X_{ij1} from normal distribution $N(0, \sigma_x^2)$ and X_{ij2} from poisson distribution with mean λ. For the random parts, we generate the random effects from $N(0, \sigma_1^2)$ for the first pN clusters and generate the last $(1-p)N$ clusters from $N(0, \sigma_2^2)$, where p is a fraction between 0 and 1. The error terms are generated from $N(0, \sigma_\varepsilon^2)$.

For each simulated data set, we fit the linear mixed effects model in R and calculate the test statistics given by (2.6) and (2.7) respectively for each cluster. The empirical distributions of the test statistics for selected clusters were shown in Figure 2.1. Under the null they approximately follow chi-square distribution while under the alternative the distributions differ from the chi-square distribution.

We varied different parameters (cluster size n_i, sample size N, variance of random effects σ_1 and σ_2, fixed effects βs, the variance of error terms σ_ε^2) and the assumed distribution parameters for X_{ij}s, and calculate the proportion of the cases for which the p-values fall below 0.05 for both the normal and abnormal clusters. Specifically, in Table 2.1 we varied different variance parameters while in Table 2.2 we considered various cluster sizes and total sample sizes when we fix the number of abnormal cluster to be 3. The size of the test was well preserved since the outlier test rejects less then 5% of all simulations for normal subjects. In general, as the difference of the variance between normal and abnormal random effects increase and as the cluster number N increases, the power of our test improves. The empirical test is slightly better than the model-based test in all cases.

We can also vary the proportion of abnormal clusters fixing the sample size and cluster size. Actually it is the same with varying cluster size fixing the number of abnormal clusters. The result is reported in Table 2.3. In table 2.4, we varied the variance of error terms σ_ε^2. We can see that as the variance of error terms

Table 2.1 The numbers in the table are the proportion of p-values fall below 0.05.

$N = 50, n_i = 10, \beta_1 = 1, \beta_2 = 3, \beta_3 = 5, \sigma_x^2 = 3, \sigma_\varepsilon^2 = 1, \lambda = 2$ and $p = 0.94$.

	Normal cluster (NH)		Abnormal cluster (AH)	
	model based	empirical	model based	empirical
$\sigma_1^2 = 1, \sigma_2^2 = 10$	0.0012	0.0016	0.6710	0.6807
$\sigma_1^2 = 1, \sigma_2^2 = 5$	0.0075	0.0100	0.5557	0.5707
$\sigma_1^2 = 3, \sigma_2^2 = 10$	0.0174	0.0189	0.4807	0.4880
$\sigma_1^2 = 1, \sigma_2^2 = 3$	0.0175	0.0238	0.3947	0.4197
$\sigma_1^2 = 5, \sigma_2^2 = 10$	0.0334	0.0359	0.2690	0.2773

Table 2.2 The numbers in the table are the proportion of p-values fall below 0.05.

$\beta_1 = 1, \beta_2 = 3, \beta_3 = 5, \sigma_1^2 = 1, \sigma_2^2 = 10, \sigma_x^2 = 3, \sigma_\varepsilon^2 = 1$ and $\lambda = 2$.

		Normal cluster (NH)		Abnormal cluster (AH)	
N	n_i	model based	empirical	model based	empirical
50	5	0.0013	0.0021	0.6503	0.6650
	10	0.0020	0.0025	0.6543	0.6600
100	5	0.0031	0.0052	0.7227	0.7350
	10	0.0035	0.0044	0.7407	0.7490
200	5	0.0070	0.0109	0.7557	0.7703
	10	0.0079	0.0099	0.7643	0.7720

Table 2.3 The numbers in the table are the proportion of p-values fall below 0.05. $N = 50, n_i = 10, \beta_1 = 1, \beta_2 = 3, \beta_3 = 5, \sigma_1^2 = 1, \sigma_2^2 = 10, \sigma_x^2 = 3, \sigma_\varepsilon^2 = 1$ and $\lambda = 2$.

	Normal cluster (NH)		Abnormal cluster (AH)	
	model based	empirical	model based	empirical
$p = 0.8$	0.0001	0.0001	0.3798	0.3862
$p = 0.9$	0.0002	0.0002	0.5550	0.5618
$p = 0.94$	0.0012	0.0016	0.6710	0.6807
$p = 0.96$	0.0038	0.0052	0.7500	0.7555
$p = 0.98$	0.0110	0.0152	0.8260	0.8420

Table 2.4 The numbers in the table are the proportion of p-values fall below 0.05. $N = 50, n_i = 10, \sigma_1^2 = 1, \sigma_2^2 = 10, \sigma_x^2 = 3, \beta_1 = 1, \beta_2 = 3, \beta_3 = 5, \lambda = 2$ and $p = 0.94$.

	Normal cluster (NH)		Abnormal cluster (AH)	
	model based	empirical	model based	empirical
$\sigma_\varepsilon^2 = 3$	0.0006	0.0040	0.5927	0.642
$\sigma_\varepsilon^2 = 2$	0.0011	0.0028	0.6323	0.6503
$\sigma_\varepsilon^2 = 1$	0.0014	0.0018	0.6650	0.6700
$\sigma_\varepsilon^2 = 0.5$	0.0017	0.0019	0.6817	0.6857

decreases, it has less impact on the model and our test performs better and better. In table 2.5, we varied fixed effects βs while in table 2.6 we consider different distribution parameters for X_{ij}s, we found that they do not have obvious impact

Table 2.5 The numbers in the table are the proportion of p-values fall below 0.05. $N = 50, n_i = 10, \sigma_1^2 = 1, \sigma_2^2 = 10, \sigma_x^2 = 1, \sigma_\varepsilon^2 = 3, \lambda = 2$ and $p = 0.94$.

	Normal cluster (NH)		Abnormal cluster (AH)	
	model based	empirical	model based	empirical
$\beta_1 = 1, \beta_2 = 3, \beta_3 = 5$	0.0014	0.0018	0.6650	0.6700
$\beta_1 = 1, \beta_2 = 5, \beta_3 = 10$	0.0014	0.0020	0.6623	0.6693
$\beta_1 = 10, \beta_2 = 5, \beta_3 = 1$	0.0018	0.0024	0.6610	0.6667
$\beta_1 = 5, \beta_2 = 5, \beta_3 = 5$	0.0016	0.0022	0.6660	0.6730
$\beta_1 = 1, \beta_2 = 10, \beta_3 = 5$	0.0014	0.0019	0.6560	0.6653

on the performance of our test.

2.5.2 Normal response with a random intercept and a random slope.

We consider a more general case that we add one more random slope term in the previous model. In this case, the 500 data sets are generated from the model

$$y_{ij} = \beta_1 + X_{ij1}\beta_2 + X_{ij2}\beta_3 + b_{i1} + X_{ij1}b_{i2} + \varepsilon_{ij}, \qquad i = 1, \cdots, N; j = 1, \cdots, n_i$$

In contrast with the random intercept case, we generate random effects for the first pN clusters from multivariate normal distribution $N(0, \Sigma_1)$ and the last $(1-p)N$

Table 2.6 The numbers in the table are the proportion of p-values fall below 0.05. $N = 50, n_i = 10, \beta_1 = 1, \beta_2 = 3, \beta_3 = 5, \sigma_1^2 = 1, \sigma_2^2 = 10, \sigma_\varepsilon^2 = 1$ and $p = 0.94$.

	Normal cluster (NH)		Abnormal cluster (AH)	
	model based	empirical	model based	empirical
$\sigma_x^2 = 3, \lambda = 2$	0.0020	0.0025	0.6647	0.6737
$\sigma_x^2 = 10, \lambda = 2$	0.0013	0.0020	0.6643	0.6720
$\sigma_x^2 = 3, \lambda = 5$	0.0013	0.0017	0.6657	0.6727
$\sigma_x^2 = 10, \lambda = 10$	0.0011	0.0014	0.6663	0.6737
$\sigma_x^2 = 3, \lambda = 10$	0.0017	0.0024	0.6513	0.6590

from $N(0, \Sigma_2)$, where

$$\Sigma_1 = \begin{pmatrix} \sigma_{11}^2 & \rho_1\sigma_{11}\sigma_{12} \\ \rho_1\sigma_{11}\sigma_{12} & \sigma_{12}^2 \end{pmatrix},$$

and

$$\Sigma_2 = \begin{pmatrix} \sigma_{21}^2 & \rho_2\sigma_{21}\sigma_{22} \\ \rho_2\sigma_{21}\sigma_{22} & \sigma_{22}^2 \end{pmatrix}.$$

The error terms ε_i are generated from $N(0, \sigma_\varepsilon^2)$. X_{ij1} are generated from $N(0, \sigma_x^2)$, X_{ij2} are generated from poisson distribution with mean λ and β's are fixed numbers as in the previous model. We repeat the simulation procedure and concluded almost the same result as in the random intercept model. Tables 2.7, Tables 2.8 and Tables 2.9 show the results that we varied variance of random effects, sample size, cluster size and proportion of abnormal clusters.

Table 2.7 The numbers in the table are the proportion of p-values fall below 0.05. $\beta_1 = 1, \beta_2 = 3, \beta_3 = 5, N = 50, n_i = 10, \sigma_x^2 = 3, \sigma_\varepsilon^2 = 1, \lambda = 2, \rho_1 = \rho_2 = 0.5$ and $p = 0.94$.

	Normal cluster (NH)		Abnormal cluster (AH)	
	model based	empirical	model based	empirical
$\sigma_{11}^2 = 1, \sigma_{12}^2 = 1, \sigma_{21}^2 = 5, \sigma_{22}^2 = 1$	0.0251	0.0314	0.2128	0.2440
$\sigma_{11}^2 = 1, \sigma_{12}^2 = 10, \sigma_{21}^2 = 10, \sigma_{22}^2 = 1$	0.0268	0.0309	0.2892	0.3148
$\sigma_{11}^2 = 1, \sigma_{12}^2 = 1, \sigma_{21}^2 = 1, \sigma_{22}^2 = 5$	0.0156	0.0248	0.3268	0.3468
$\sigma_{11}^2 = 1, \sigma_{12}^2 = 1, \sigma_{21}^2 = 5, \sigma_{22}^2 = 5$	0.0141	0.0188	0.3880	0.4140
$\sigma_{11}^2 = 1, \sigma_{12}^2 = 1, \sigma_{21}^2 = 10, \sigma_{22}^2 = 10$	0.0073	0.0101	0.5392	0.5668

Similarly, we consider various variance of error terms σ_ε^2, fixed effects βs and the assumed distribution parameters for X_{ij}s. The results are reported respectively in Tables 2.10, Tables 2.11 and Tables 2.12.

2.5.3 Poisson response with a random intercept

We conduct a simulation study for the poisson response under GLMM mentioned above. The model is given by

$$\log \mu_{ij} = \beta_1 + X_{ij1}\beta_2 + X_{ij2}\beta_3 + b_i, \qquad i = 1, \cdots, N; j = 1, \cdots, n_i$$

Table 2.8 The numbers in the table are the proportion of p-values fall below 0.05 with 3 abnormal cluster. $\beta_1 = 1, \beta_2 = 3, \beta_3 = 5, \sigma^2_{11} = \sigma^2_{12} = 1, \sigma^2_{21} = \sigma^2_{22} = 10, \sigma^2_x = 3, \sigma^2_\varepsilon = 1, \lambda = 2$ and $\rho_1 = \rho_2 = 0.5$.

		Normal cluster (NH)		Abnormal cluster (AH)	
N	n_i	model based	empirical	model based	empirical
50	5	0.0095	0.0193	0.5613	0.6217
	10	0.0129	0.0184	0.6060	0.6287
100	5	0.0162	0.0297	0.6207	0.6597
	10	0.0205	0.0276	0.6517	0.6710
200	5	0.0212	0.0453	0.6327	0.6667
	10	0.0549	0.0712	0.6760	0.6957

where the covariates and random effects are generated from the same specification in previous sections and the outcome observations are drawn from Poisson(μ_{ij}). We fit the poisson mixed model and apply both the model based and empirical tests to find the proportion of p-values fall below 0.05 in each case varying the variance of random effects as well as sample size and cluster size. The results are summarized in the Tables 2.13 and 2.14.

Alternatively, we consider the GLMM with overdispersion. We apply the additive overdispersion model and the model is given by

$$\log \mu_{ij} = \beta_1 + X_{ij1}\beta_2 + X_{ij2}\beta_3 + b_i + \varepsilon_{ij}, \qquad i = 1, \cdots, N; j = 1, \cdots, n_i$$

Table 2.9 The numbers in the table are the proportion of p-values fall below 0.05. $N = 50, n_i = 10, \beta_1 = 1, \beta_2 = 3, \beta_3 = 5, \sigma_{11}^2 = \sigma_{12}^2 = 1, \sigma_{21}^2 = \sigma_{22}^2 = 10, \sigma_x^2 = 3, \sigma_\varepsilon^2 = 1, \lambda = 2$ and $\rho_1 = \rho_2 = 0.5$.

	Normal cluster (NH)		Abnormal cluster (AH)	
	model based	empirical	model based	empirical
$p = 0.8$	0.0015	0.0024	0.4052	0.4274
$p = 0.9$	0.0072	0.0098	0.5228	0.5504
$p = 0.94$	0.0131	0.0190	0.5980	0.6267
$p = 0.96$	0.0173	0.0256	0.6320	0.6590
$p = 0.98$	0.0222	0.0324	0.6540	0.6740

Table 2.10 The numbers in the table are the proportion of p-values fall below 0.05. $\beta_1 = 1, \beta_2 = 3, \beta_3 = 5, N = 50, n_i = 10, \sigma_x^2 = 3, \sigma_{11}^2 = \sigma_{12}^2 = 1, \sigma_{21}^2 = \sigma_{22}^2 = 10, \lambda = 2, \rho_1 = \rho_2 = 0.5$ and $p = 0.98$.

	Normal cluster (NH)		Abnormal cluster (AH)	
	model based	empirical	model based	empirical
$\sigma_\varepsilon^2 = 3$	0.0075	0.0366	0.4900	0.5940
$\sigma_\varepsilon^2 = 2$	0.0138	0.0368	0.5820	0.6780
$\sigma_\varepsilon^2 = 1$	0.0250	0.0342	0.6580	0.6740
$\sigma_\varepsilon^2 = 0.5$	0.0301	0.0343	0.7100	0.7220

Table 2.11 The numbers in the table are the proportion of p-values fall below 0.05. $N = 50, n_i = 10, \sigma_x^2 = 3, \sigma_{11}^2 = \sigma_{12}^2 = 1, \sigma_{21}^2 = \sigma_{22}^2 = 10, \sigma_\varepsilon^2 = 1, \lambda = 2, \rho_1 = \rho_2 = 0.5$ and $p = 0.98$.

	Normal cluster (NH)		Abnormal cluster (AH)	
	model based	empirical	model based	empirical
$\beta_1 = 1, \beta_2 = 3, \beta_3 = 5$	0.0245	0.0349	0.6440	0.6700
$\beta_1 = 1, \beta_2 = 5, \beta_3 = 10$	0.0221	0.0341	0.6900	0.7200
$\beta_1 = 10, \beta_2 = 5, \beta_3 = 1$	0.0233	0.0331	0.6920	0.7220
$\beta_1 = 5, \beta_2 = 5, \beta_3 = 5$	0.0248	0.0333	0.6620	0.6920
$\beta_1 = 1, \beta_2 = 10, \beta_3 = 5$	0.0242	0.0338	0.6560	0.6920

Table 2.12 The numbers in the table are the proportion of p-values fall below 0.05. $\beta_1 = 1, \beta_2 = 3, \beta_3 = 5, N = 50, n_i = 10, \sigma_{11}^2 = \sigma_{12}^2 = 1, \sigma_{21}^2 = \sigma_{22}^2 = 10, \sigma_\varepsilon^2 = 1, \rho_1 = \rho_2 = 0.5$ and $p = 0.98$.

	Normal cluster (NH)		Abnormal cluster (AH)	
	model based	empirical	model based	empirical
$\sigma_x^2 = 3, \lambda = 2$	0.0250	0.0342	0.6580	0.6740
$\sigma_x^2 = 10, \lambda = 2$	0.0242	0.0344	0.6820	0.7060
$\sigma_x^2 = 3, \lambda = 5$	0.0242	0.0339	0.6840	0.7080
$\sigma_x^2 = 10, \lambda = 10$	0.0230	0.0327	0.6960	0.7140
$\sigma_x^2 = 3, \lambda = 10$	0.0250	0.0360	0.6920	0.7180

Table 2.13 The numbers in the table are the proportion of p-values fall below 0.05.

$\beta_1 = 3, \beta_2 = 2, \beta_3 = -1, N = 50, n_i = 10, \sigma_x^2 = 3, \lambda = 3, p = 0.98$

	Normal cluster (NH)		Abnormal cluster (AH)	
	model based	empirical	model based	empirical
$\sigma_1^2 = 1, \sigma_2^2 = 10$	0.0179	0.0173	0.8936	0.8936
$\sigma_1^2 = 1, \sigma_2^2 = 5$	0.0144	0.0141	0.6778	0.6778
$\sigma_1^2 = 2, \sigma_2^2 = 10$	0.0122	0.0122	0.5979	0.5979
$\sigma_1^2 = 1, \sigma_2^2 = 3$	0.0092	0.0092	0.4500	0.4500
$\sigma_1^2 = 2, \sigma_2^2 = 5$	0.0082	0.0082	0.4027	0.4027

Table 2.14 The numbers in the table are the proportion of p-values fall below 0.05 with 3 abnormal clusters in each case. $\beta_1 = 3, \beta_2 = 2, \beta_3 = -1, \sigma_1^2 = 1, \sigma_2^2 = 10, \sigma_x^2 = 2, \lambda = 3$.

		Normal cluster(NH)		Abnormal cluster(AH)	
N	n_i	model based	empirical	model based	empirical
50	5	0.0057	0.0092	0.3493	0.4207
	10	0.0052	0.0074	0.3907	0.4560
100	5	0.0094	0.0133	0.3860	0.4847
	10	0.0093	0.0111	0.5173	0.5713
200	5	0.0123	0.0163	0.4013	0.5033
	10	0.0175	0.0195	0.5233	0.5853

Table 2.15 The numbers in the table are the proportion of p-values fall below 0.05. $\beta_1 = 3, \beta_2 = 2, \beta_3 = -1, N = 50, n_i = 10, \sigma_x^2 = 3, \sigma_\varepsilon^2 = 0.1, \lambda = 3$ and $p = 0.98$

	Normal cluster (NH)		Abnormal cluster (AH)	
	model based	empirical	model based	empirical
$\sigma_1^2 = 1, \sigma_2^2 = 10$	0.0026	0.0028	0.3880	0.3920
$\sigma_1^2 = 1, \sigma_2^2 = 5$	0.0027	0.0028	0.2820	0.2900
$\sigma_1^2 = 3, \sigma_2^2 = 10$	0.0010	0.0011	0.3200	0.3200
$\sigma_1^2 = 1, \sigma_2^2 = 3$	0.0033	0.0034	0.1800	0.1880

where ε_{ij}s are generated from normal distribution $N(0, \sigma_\varepsilon^2)$. In this case, the estimated variance in (2.6) and (2.7) will be adjusted by multiplying the estimated dispersion parameter. The results are summarized in the Tables 2.15 and 2.16.

The results are similar to those for the linear mixed models. As the ratio of the variance of the normal and abnormal clusters decreases and as the number of clusters increases given the number of abnormal clusters fixed, our test statistic performs better under the alternative hypothesis.

Table 2.16 The numbers in the table are the proportion of p-values fall below 0.05. $\beta_1 = 3, \beta_2 = 2, \beta_3 = -1, \sigma_1^2 = 1, \sigma_2^2 = 10, \sigma_x^2 = 2, \sigma_\varepsilon^2 = 0.1, \lambda = 3$ given there is one abnormal cluster.

		Normal cluster(NH)		Abnormal cluster(AH)	
N	n_i	model based	empirical	model based	empirical
50	5	0.0135	0.0156	0.3700	0.3760
	10	0.0038	0.0038	0.3100	0.3120
100	5	0.0109	0.0131	0.4080	0.4240
	10	0.0029	0.0030	0.3680	0.3760
200	5	0.0084	0.0103	0.3980	0.4140
	10	0.0031	0.0033	0.4040	0.4100

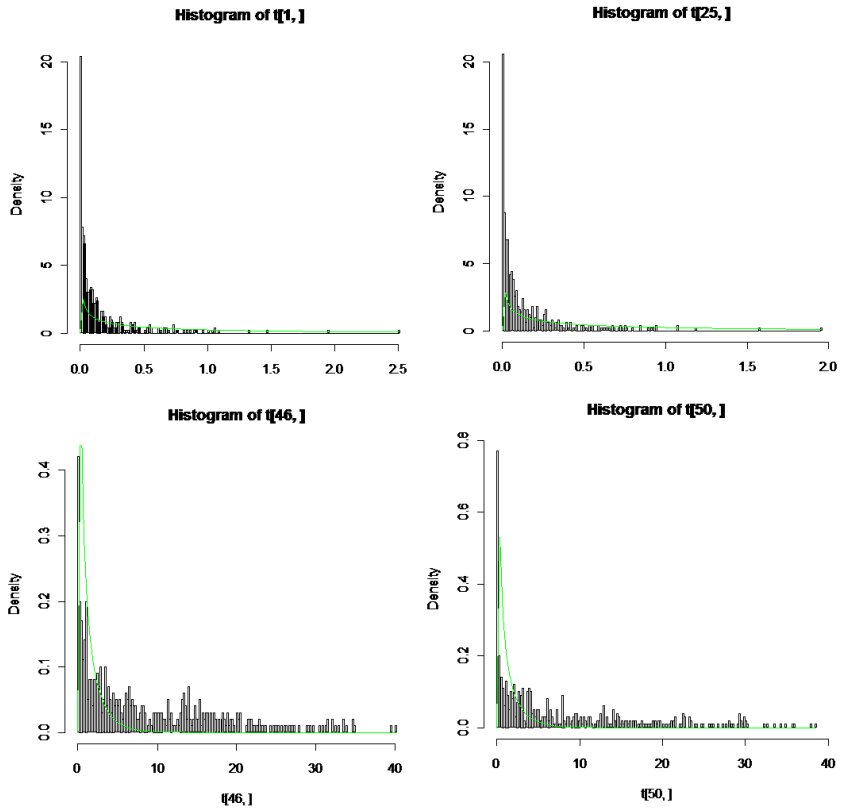

Figure 2.1 The histogram plot of 500 test statistics along with the graph of chi-square distribution with degree of freedom 1. The two plots in the top are for clusters 1 and 25 from the normal random effects clusters, and the two plots in the bottom are for clusters 46 and 50 from the abnormal random effects clusters.

CHAPTER 3

Case study

3.1 Background and data information

Rheumatoid arthritis (RA) is a chronic and progressive disease that leads to inflammation of the joints and surrounding tissues. Wrists, fingers, knees, feet, and ankles are the most commonly affected. The disease often begins slowly, usually with only minor joint pain, stiffness, and fatigue. The cause of RA is unknown. It is an autoimmune disease, which means the body's immune system mistakenly attacks healthy tissue. RA can occur at any age, but is more common in middle age. There is no test that can determine for sure whether someone has RA, since

for some patients all test results will be normal. Two lab tests that often help in the diagnosis are Rheumatoid factor test and Anti-CCP antibody test (Klareskog et al. (2009)). For the treatment for RA, other than simply medications, biologic agents are used as well. They are designed to affect specific parts of the immune system that play a role in the disease process of rheumatoid arthritis. Over the past decade, enhanced understanding of the molecular pathogenesis has led to the development of biologic agents. These innovative treatments have changed the outcomes for patients and face of RA. Many research addressed on biologic agents for their treatment comparison and efficacy estimation.

We apply our test to a longitudinal data set in which 808 RA patients who were followed for three years. Clinical and demographic data were collected at at baseline, then at the first, the second and the third years. The number of repeated measurements is 4 over follow-up for each individual. Baseline covariates include age, gender, and duration of disease (in months). A blood sample was taken from all patients for anti-cyclic citrullinated peptide antibody (anti-CCP) and C-Reactive Protein (CRP). The Health Assessment Questionaire (HAQ) score were repeatedly measured at every visit, which is based on a self-reported questionnaire. In addition, the swollen joint count, the tender joint count (out of 28 defined joints) and the count of both swollen and tender (both28) are measured. Treatment status is recorded during follow-up as an indicator term (treatment), being one if the

patient was on any treatment.

3.2 Statistical model

3.2.1 Separate Models

We fitted a linear mixed model to the repeated HAQ data by including 6 fixed effects associated with covariates antiCCP, CRP, treatment, gender, age, and disease duration. In this case, we assume one random intercept term for each individual. We then calculated the test statistic and obtained the p-value for each individuals. For the individuals with p-value smaller than the significant level 0.05, the hypothesis that the random effect are drawn from the assumed normal distribution with a constant variance is rejected.

In the real data set, with the model-based test there are 29 out of 808 people with p-value smaller than 0.05. We then considered removing them from the data set and fitting the linear mixed model again. The same procedure can be applied using the empirical test which gave 36 abnormal individuals. These 36 subjects include all the 29 abnormal subjects identified by the model based method. The refitted results are summarized in Table 3.1. While the significance of all the coefficients remain the same, the coefficients for disease duration (duration) are

Table 3.1 The fitted results for 6 fixed effects and the variance for random effects for full data set and the adjusted data set using both the model based method and empirical method for the continuous response HAQ in linear mixed model.

	Full data			Adjusted data (model based)			Adjusted data (empirical)		
	coef	s.e	p-value	coef	s.e	p-value	coef	s.e	p-value
Intercept	0.4281	0.0846	< 0.0001	0.3821	0.0780	< 0.0001	0.3796	0.0766	< 0.0001
antiCCP	0.1702	0.0493	0.0006	0.2136	0.0448	< 0.0001	0.2091	0.0440	< 0.0001
CRP	0.0036	0.0004	< 0.0001	0.0037	0.0004	< 0.0001	0.0038	0.0004	< 0.0001
treatment	0.1032	0.0259	< 0.0001	0.1136	0.0260	< 0.0001	0.1189	0.0261	< 0.0001
gender	-0.3147	0.0468	< 0.0001	-0.3365	0.0427	< 0.0001	-0.3386	0.0418	< 0.0001
age	0.0079	0.0015	< 0.0001	0.0031	0.0014	< 0.0001	0.0075	0.0013	< 0.0001
duraton	-0.0012	0.0006	0.0352	-0.0015	0.0006	0.0079	-0.0015	0.0006	0.0072
σ^2	0.34939			0.26996			0.25437		

less significant after removing the outliers. The absolute values of the coefficients appear slightly larger for most covariates in refitted models. The variance estimates for random effects shrink by almost 25% after the outliers are taken out.

Similarly, we can fit a poisson mixed effects model for the count outcome both28. The random and fixed effects are specified in the same way as in the linear mixed model. In this case, the model based test identified 7 abnormal individuals while the empirical test statistics identified 23 abnormal subjects. The 7 abnormal cases from the model-based method are a proper subset of the 23 from the empirical method. The refitted results for the poisson mixed model after deleting the outliters are summarized in Table 3.2. The results are similar to those

Table 3.2 The fitted results for 6 fixed effects and the variance for random effects for full data set and the adjusted data set using both the model based method and empirical method for the count response both28 in poisson mixed model.

	Full data			Adjusted data (model based)			Adjusted data (empirical)		
	coef	s.e	p-value	coef	s.e	p-value	coef	s.e	p-value
Intercept	0.8394	0.1746	< 0.0001	0.8347	0.1721	< 0.0001	0.7956	0.1684	< 0.0001
antiCCP	0.4459	0.1003	< 0.0001	0.4770	0.0985	< 0.0001	0.5188	0.0959	< 0.0001
CRP	0.0042	0.0005	< 0.0001	0.0050	0.0005	< 0.0001	0.0050	0.0005	< 0.0001
treatment	0.0825	0.0349	0.0181	0.1043	0.0358	0.0035	0.1181	0.0380	0.0019
gender	-0.5951	0.0985	< 0.0001	-0.6300	0.09718	< 0.0001	-0.6411	0.0946	< 0.0001
age	0.0033	0.0031	0.2825	0.0031	0.0030	0.3103	0.0037	0.0029	0.2088
duration	-0.0260	0.0009	< 0.0001	-0.0269	0.0009	< 0.0001	-0.0285	0.0009	0.0072
σ^2	1.4629			1.39520			1.28390		

for HAQ response with significance unchanged and magnitude of the coefficients slightly increased.

3.2.2 Joint Model

Since the two responses HAQ and $both28$ may be dependent, we then consider fitting the linear mixed model and poisson mixed model jointly, assuming the random effects to be generated from a bivariate normal distribution with a 2×2 covariance matrix. This type of joint modelling was carried out in SAS using `PROC NLMIXED`. We then perform outlier test for all the subjects again from the fitted joint model. Under the model based test, 15 abnormal subjects are removed and

under the empirical test, 28 subjects are removed. The 28 abnormal subjects from the empirical method include all the 15 abnormal subjects we found in model based method. The estimated results for both the full data and the adjusted data are summarized in Table 3.3. As in the separated models, the estimate of the variance of random effects decreases after adjusting the data by taking out the outliers. In the joint model, the correlation coefficient for the two random effects increases after deleting abnormal subjects.

For the joint model, we found that the p-value for the coefficient of *treatment* in the poisson response part decreases from 0.0621 to 0.0115 and 0.0074 when we delete abnormal individuals applying the model based method and empirical method, respectively. The coefficient becomes significant after adjusting the data, suggesting that our method may potentially change the interpretation of the results.

We compare the outlier identification results between the separate models and the joint model for both the model based and empirical methods. The results are summarized in Figures 3.1 and 3.2. Applying the model based method, there are 3 common abnormal subjects for both the linear mixed model and poisson mixed model, and 2 of them belongs to the abnormal group we found from the joint model. Among the 15 abnormal subjects in the joint model, 11 are from the abnormal group in linear mixed model and only 2 are not from the abnormal

Table 3.3 The fitted results of the joint model of the continuous response HAQ and poisson response both28 for full data set and the adjusted data set using both the model based test and the empirical test. σ_1^2 and σ_2^2 are the variances of the random intercepts for HAQ and both28, respectively; ρ is the correlation between the two random effects.

	Full data			Adjusted data (model based)			Adjusted data (empirical)		
HAQ	coef	s.e	p-value	coef	s.e	p-value	coef	s.e	p-value
intercept	0.4382	0.0845	< 0.0001	0.4320	0.0805	< 0.0001	0.4065	0.0803	< 0.0001
antiCCP	0.1717	0.0491	0.0005	0.1725	0.0468	0.0002	0.1810	0.0464	0.0001
CRP	0.0036	0.0004	< 0.0001	0.0037	0.0004	< 0.0001	0.0037	0.00004	< 0.0001
treatment	0.0981	0.0254	0.0001	0.1146	0.0254	< 0.0001	0.1089	0.0256	< 0.0001
gender	-0.3152	0.0468	< 0.0001	-0.3345	0.0445	< 0.0001	-0.3395	0.0443	< 0.0001
age	0.0080	0.0015	< 0.0001	0.0076	0.0014	< 0.0001	0.0080	0.0014	< 0.0001
duration	-0.0016	0.0005	0.0041	-0.0017	0.0006	0.0023	-0.0018	0.0006	0.0018
both28	coef	s.e	p-value	coef	s.e	p-value	coef	s.e	p-value
intercept	0.8633	0.1736	< 0.0001	0.8603	0.1733	< 0.0001	0.8228	0.1720	< 0.0001
antiCCP	0.4329	0.0998	< 0.0001	0.4497	0.0996	< 0.0001	0.5020	0.0981	< 0.0001
CRP	0.0039	0.0005	< 0.0001	0.0050	0.0005	< 0.0001	0.0051	0.0005	< 0.0001
treatment	0.0644	0.0345	0.0621	0.0892	0.0352	0.0115	0.0983	0.0366	0.0074
gender	-0.5706	0.0979	< 0.0001	-0.5752	0.0978	< 0.0001	-0.5979	0.0970	< 0.0001
age	0.0030	0.0031	0.3308	0.3231	0.0031	0.4441	0.0025	0.0030	0.4096
duration	-0.0258	0.0009	< 0.0001	-0.0260	0.0009	< 0.0001	-0.0268	0.0009	< 0.0001
σ_1^2	0.3481			0.3036			0.2942		
σ_2^2	1.4422			1.4024			1.3412		
ρ	0.5367			0.5415			0.5524		

group in the separated models. Applying empirical method, there are 8 common abnormal subjects for both the linear mixed model and poisson mixed model, and they all belong to the abnormal group we found from the joint model. Among the

28 abnormal subjects in the joint model, 18 are from the abnormal group in linear mixed model and 11 are from the poisson mixed model, and 7 not belong to any abnormal group from the separated models.

It seems that that the separated model may overstate or understate the number of abnormal individuals. The joint model computes a moderate amount of abnormal people contains almost all the abnormal people appears in both separated models and the joint model uses more information. The joint model may be a better way to find the realistic result.

3.2.3 Random slope

Next we try to fit the mixed models with the same structure but an additional random slope for the disease duration. For the HAQ outcome, we fitted linear mixed effects model and performed both the model based test and the empirical test, identifying 14 outliers under the model based test and 44 outliers under the empirical test. The refitted results are then summarized in Table 3.4. The interpretation for the results is similar to the previous analysis.

Similarly, we fitted a poisson mixed effects model with random intercept and slope for the count outcome both28. Under the model based test, 5 abnormal subjects are removed and under the empirical test, 32 subjects are removed. The

Table 3.4 The fitted results for 6 fixed effects and the variance for random effects for full data set and the adjusted data set using both the model based method and empirical method for HAQ under linear mixed model random intercept and slope.

	Full data			Adjusted data (model based)			Adjusted data (empirical)		
	coef	s.e	p-value	coef	s.e	p-value	coef	s.e	p-value
Intercept	0.4410	0.0826	< 0.0001	0.4279	0.0796	< 0.0001	0.4109	0.0779	< 0.0001
antiCCP	0.1719	0.0482	< 0.0001	0.1779	0.0463	< 0.0001	0.2057	0.0455	< 0.0001
CRP	0.0033	0.0004	< 0.0001	0.0033	0.0004	< 0.0001	0.0036	0.0004	< 0.0001
treatment	0.0691	0.0259	< 0.0001	0.0876	0.0257	< 0.0001	0.0798	0.0254	0.0019
gender	-0.3041	0.0460	< 0.0001	-0.3252	0.0442	< 0.0001	-0.3502	0.0433	< 0.0001
age	0.0077	0.0015	< 0.0001	0.0076	0.0014	< 0.0001	0.0073	0.0013	< 0.0001
duration	0.0012	0.0007	0.0512	-0.0016	0.0007	0.1329	-0.0013	0.0006	0.0325
σ^2_{INT}	0.3319			0.2984			0.2055		
σ^2_{SLP}	0.0002			0.0001			0.0001		

refitted results for the poisson mixed model are summarized in Table 3.5. The interpretation for the results is similar to the previous analysis.

Table 3.5 The fitted results for 6 fixed effects and the variance for random effects for full data set and the adjusted data set using both the model based method and empirical method for both28 under poisson mixed model with random intercept and slope.

	Full data			Adjusted data (model based)			Adjusted data (empirical)		
	coef	s.e	p-value	coef	s.e	p-value	coef	s.e	p-value
Intercept	0.9541	0.1910	< 0.0001	0.9837	0.1869	< 0.0001	0.8938	0.1840	< 0.0001
antiCCP	0.4174	0.1075	0.0001	0.4443	0.1052	< 0.0001	0.4932	0.1038	< 0.0001
crp	0.0027	0.0006	< 0.0001	0.0028	0.0006	< 0.0001	0.0028	0.0006	< 0.0001
treatment	0.0982	0.0471	0.0369	0.1026	0.0471	0.0292	0.0848	0.0477	0.0754
gender	-0.6195	0.1056	< 0.0001	-0.6441	0.1038	< 0.0001	-0.6981	0.1031	< 0.0001
ageassess	0.0038	0.0033	0.2516	0.0029	0.0033	0.3778	0.0041	0.0032	0.1965
disdur	-0.0445	0.0023	< 0.0001	-0.0439	0.0023	< 0.0001	-0.0429	0.0021	< 0.0001
σ^2_{INT}	2.3504			2.1969			1.8848		
σ^2_{SLP}	0.0027			0.0026			0.0020		

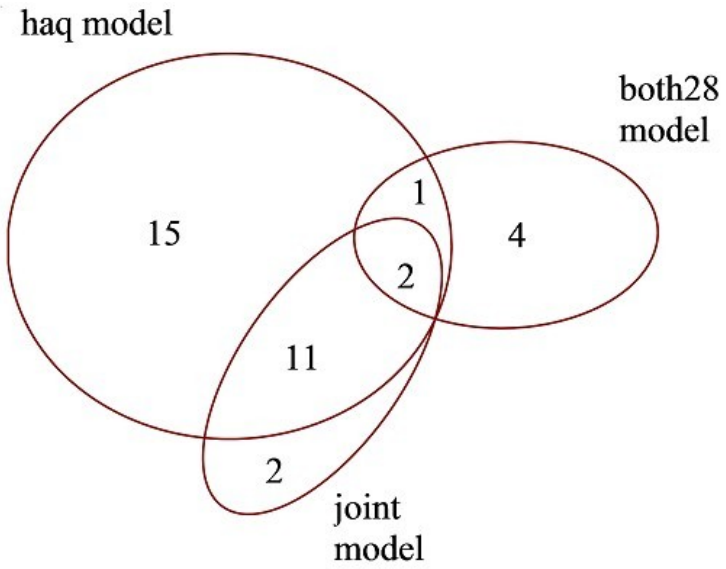

Figure 3.1 Venn diagram for the sets of abnormal subjects identified for three models by using the model based test. The numbers are the size of the sets.

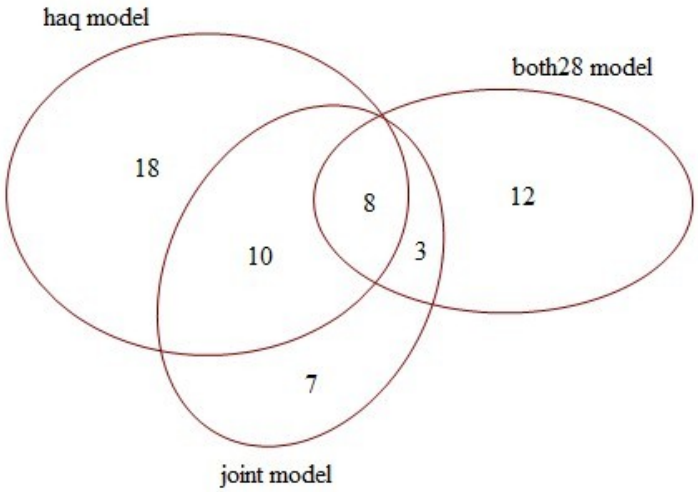

Figure 3.2 Venn diagram for the sets of abnormal subjects identified for three models by using the empirical test. The numbers are the size of the sets.

CHAPTER 4

Discussion

We are interested in identifying clusters in longitudinal data analysis that violates the equal variance assumption. In the application to a rheumatoid arthritis cohort study, we find that this kind of heterogeneity in the data is likely to induce bias in estimating the impact of important risk factors. Conventional analysis may underestimate their impact in absolute scale when the homogeneity assumption is violated. This thesis focuses on testing the homogeneity assumption in mixed-effects model and the proposed tests are useful for model diagnostic checking in the analysis of longitudinal data. Removing "outliers" is a straightforward solution, but we need to be more careful in the real data analysis. When the number of

clusters is moderate or small, investigators may not want to remove one whole cluster in order to preserve sufficient sample size for the statistical analysis. In that case, individually examining which observation from the abnormal cluster could refine the model checking procedure. One may then choose to only remove specific abnormal observations from such clusters.

Besides generalized linear models, longitudinal data are also frequently analyzed by nonparametric and semi-parametric models. For example, varying coefficient models are important tool to explore the dynamic pattern in many scientific areas and becoming more and more attractive to both applied and methodological statisticians (Fan and Zhang (2008)). The varying coefficient models considering the dynamic feature which may exist in the data set are firstly introduced by Cleveland, Grosse and Shyu (1991). This semi-parametric technique allows the coefficients to vary smoothly over the group and permits nonlinear interactions. Varying coefficient models can be extended to varying coefficient mixed models by adding the random effects term. To check the model assumption of equal variance in the models, one may follow a similar paradigm by computing the empirical estimator of the random effects as well as their covariance matrix and calculating the test statistic given in this thesis. Other nonparametric and semi-parametric models include nonparametric mixed effects models (Wang (1998), Guo (2002)), generalized additive mixed models (Hastie and Tibshirani (1990), Lin and Zhang

(1999)), partially linear mixed models (Wahba (1984), Green and Silverman (1994)) and semi-parametric Threshold Model (Tong (1990), Li and Zhang (2011), among others). The model-based covariance matrix may be rather complicated in those situations but an empirical estimator can always be easily computed. Further research of extenting our tests to nonparametric and semi-parametric models is needed.

Bibliography

[1] Albert P. (2008). Modeling longitudinal biomarker data with multiple assays which have known detection limits. *Biometrics* **64**, 527-537.

[2] Benjamin, R.S. and Amy, H.H. (2009). Testing random effects in the linear mixed model using approximate Bayes factors. *Biometrics* **65**, 369-376.

[3] Berridge, D.M. and Crouchley, R. (2011). *Multivariate generalized linear mixed models using R,* CRC Press.

[4] Booth, J.G. and Hobert, J.P. (2009). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B* **65**, 265-285.

[5] Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9-25.

[6] Browne, W.J., Subramanian, S.V., Jones, K. and Goldstein, H. (2005) Variance partitioning in multilevel logistic models that exhibit overdispersion. *Journal of the Royal Statistical Society* **168**, 599-613.

[7] Commenges, D. and Jacqmin-Gadda, H. (1997). Generalized score test of homogeneity based on correlated random effects models. *Journal of the Royal Statistical Society: Series B* **59**, 157-171.

[8] Cox, D.R. and Hinkeley, D.V. (1974) *Theoretical Statistics,* Chapman & Hall, London.

[9] Cleveland, W.S., Grosse, E. and Shyu, W.M. (1991) *Local regression models, in Statistical Models in S,* 309-376 Chapman & Hall, New York.

[10] Crainiceanu, C.M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B* **66**, 165-185.

[11] Demidenko, E. (2004) *Mixed models–theory and applications,* Wiley, New York.

[12] Diggle, P., Heagerty, P., Liang, K.Y. and Zeger, S. (2002). *Analysis of Longitudinal data,* Oxford University Press.

[13] Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and Its Interface* **1**, 179-195.

[14] Goldstein, P. (2003). *Multilevel Statistical Models,* 3rd ed. London: Edward Arnold.

[15] Gordon, P., West, J., Jones, H. and Gibson, T (2001). A 10 year prospective followup of patients with rheumatoid arthritis 1986-96. *Journal of Rheumatology* **28**, 2409-2415.

[16] Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach,* Chapman & Hall, London.

[17] Guo, W. (2002). Functional mixed effects models. *Biometrics* **58**, 121-128.

[18] Harville, D.A. (1997). *Matrix algebra: exercises amd solutions,* New York: Springer-Verlag.

[19] Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized additive models,* Chapman & Hall, London.

[20] Hinde, J. and Demetrio, C.G. (1998). Overdispersion: models and estimation. *Computational Statistics & Data Analysis* **27**, 151-170.

[21] Huang, X. (2009). Diagnosis of random-effect model misspecification in generalized linear mixed models for binary response. *Biometrics* **65**, 361-368.

[22] Klareskog, L., Catrina, A and Paget, S. (2009). Rheumatoid arthritis. *The lancet* **979**, 659-672.

[23] Kirwan, J.R. and Reeback, J.S. (1986). Stanford Health Assessment Questionnaire modified to assess disability in British patients with rheumatoid arthritis. *British Journal of Rheumatology* **25**, 206-209.

[24] Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-974.

[25] Li, J., Gray, B.R. and Bates, D.M. (2008). An empirical study of statistical properties of variance partition coefficients for multi-level logistic regression models. *Communications in Statistics-Simulation and Computation* **37**, 2010-2026.

[26] Li, J. and Zhang, W. (2011). A Semiparametric threshold model for censored longitudinal data analysis. *Journal of the American Statistical Association* **106**, 685-696.

[27] Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the Royal Statistical Society: Series B* **61**, 381-400.

[28] Magnus, J.R. (1988). *Linear Structures,* London: Oxfored University Press.

[29] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models,* 2nd ed. Chapman & Hall, London.

[30] McCullagh, C.E. (1989). Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association* **92**, 162-170.

[31] Molenberghs, G. and Verbeke, G. (2007). Likelihoodd ratio, score, and Wald tests in a constrained parameter space. *Journal of the American Statistical Association* **61**, 22-27.

[32] Molenberghs, G., Verbeke, G. and Demetrio, C.G. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime data analysis* **13**, 513-531.

[33] Nelder, J.A and Wedderburn, R.W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A*, 370-384.

[34] Neuhaus, J., McCulloch, C. and Boylan, R. (2011). A note on type II error under random effects misspecification in generalized linear mixed models. *Biometrics* **67**, 654-660.

[35] Pinheiro, J.C. (1994). *Topics in mixed-effects models,* Ph.D thesis, University of Wisconsin, Madison, WI.

[36] Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-effects models in S and S-PLUS,* New York: Springer-Verlag.

[37] Pugner, K.M., Scott, D.I., Holmes, J.W. and Hieke, K. (2000). The costs of rheumatoid arthritis: an international long-term view. *Seminars in arthritis and rheumatism* **29**, 305-320.

[38] Rao, C.R. and Toutenburg, H. (1999). *Linear models: least squares and alternatives.* New York: Springer-Verlag.

[39] Robinson, G.K. (1991). That BLUP is a good thing: estimation of random effects. *Statistical Science* **6**, 15-51.

[40] Rosenbaum, P.R. (2002). *Observational Studies,* New York: Springer-Verlag.

[41] Schoenbach, V.J. and Wayne, D.R. (2000). *Understanding the fundamentals of epidemiology: an evolving text,* Chapel Hill: North Carolina.

[42] Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and the likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605-610.

[43] Silvapulle, M.J. (1992). Robust Wald-type tests of one-sided hypotheses in the linear model. *Journal of the American Statistical Association* **87**, 156-161.

[44] Snijders, T. and Bosker, R. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling.* London: Sage.

[45] Stram, D.O. and Lee, J.W. (1994). Variance components testing in the longitudinal mixed effets model. *Biometrics* **50**, 254-262.

[46] Symmons, D., Barrett, E., Bankhead, C., Scott, D. and Silman, A. (1994). The incidence of rheumatoid arthritis in the United Kingdom: results for the Norfolk arthritis register. *British Journal of Rheumatology* **33**, 735-739.

[47] Tanaka, E., Mannalithara, A., Inoue, E., Hara, M., Tomatsu, T. and Kamatani, N. (2008). Efficient management of rheumatoid arthritis significantly reduces long-term functional disability. *Annals of Rheumatic Diseases* **67**, 1153-1158.

[48] Tong, H. (1990). *Non-linear time series: a dynamical system approach,* London: Oxfored University Press.

[49] Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data,* New York: Springer-Verlag.

[50] Verbeke, G. and Molenberghs, G. (2003). The us of score tests for inference on variance components. *Biometrics* **59**, 254-262.

[51] Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society: Series B* **60**, 159-174.

[52] Wahba, G. (1984). Partial spline models for the semiparametric estimation of functions of several variables. In *Statistical Analyses for Time Series,* Japan-US Joint Seminar 319-329. Institute of Statistical Mathematics, Tokyo.

[53] Wolfe, F., Michaud, K., Gefeller, O. and Choi, H.K. (2003). Redicting mortality in patients with rheumatoid arthritis. *Arthritis and Rheumatism* **48**, 1530-1542.

[54] Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79-86.